NOVA: Sparse Control, Dense Synthesis for Pair-Free Video Editing

Supplementary Material

A. Details of Experiments on Naive Multi-Keyframe Guidance

To explore the feasibility of multi-keyframe guidance for temporally consistent video editing, we conduct preliminary experiments using the reference-to-video diffusion model WAN VACE. VACE accepts two auxiliary inputs: (1) a *reference video*, which provides semantic or structural guidance (e.g., edited RGB frames, depth maps, or pose skeletons), and (2) a *mask video*, which indicates per-frame regions to be preserved (black pixels) or regenerated (white pixels).

Our initial attempt adopts a straightforward multi-keyframe strategy: we construct the reference video by placing the user-edited keyframes at their corresponding timestamps, while filling all non-keyframe positions with a uniform gray frame (RGB value 127). Concurrently, the mask video is set to black at keyframes (instructing the model to faithfully reconstruct the edited content) and white elsewhere (prompting content generation based on the reference). However, this naive setup leads to severe temporal flickering and inconsistent motion in both the 1.3B and 14B parameter variants of VACE. We hypothesize that this instability arises because VACE was not trained on such sparse, discontinuous reference signals, resulting in hallucinated textures and incoherent motion interpolation between keyframes.

To mitigate this issue, we design a more stable reference configuration. Specifically, we retain the edited RGB image only at the first keyframe (typically the first frame of the video) in the reference video. For all subsequent keyframes, we replace the edited RGB content with the corresponding *depth maps* derived from the original edited frames. Depth maps are chosen because they encode structural and motion cues with significantly less high-frequency appearance variation than RGB images, thereby reducing the risk of introducing inconsistent visual signals. Meanwhile, the mask video is simplified: only the first frame is set to black (to anchor the edited appearance), and all remaining frames (including other keyframes—are set to white), indicating that the model should synthesize these frames using the reference guidance.

This refined setup effectively decouples appearance and motion guidance: the model learns the target appearance from the first edited frame, while leveraging depth-derived structural cues from later keyframes to preserve the original video's motion dynamics. Under this configuration, VACE produces visibly more stable outputs with reduced flickering, enabling us to meaningfully analyze the limitations of sparse multi-keyframe editing

B. Details of Training Pipelines

Algorithm 1 Source Fidelity Pipeline (Pseudo-Source Generation)

Require: Target video $\mathcal{X} = \{\boldsymbol{x}_t\}_{t=0}^{T}$, Filler video pool $\mathcal{P}$
Ensure: Pseudo-source video $\tilde{\mathcal{X}} = \{\tilde{\boldsymbol{x}}_t\}_{t=0}^{T}$
Sample filler video $\mathcal{Y} = \{\boldsymbol{y}_t\} \sim \mathcal{P}$
Init mask shape $S \in \{\text{rect, ellipse, poly}\}$
Init position $\boldsymbol{p}_0$, motion $\boldsymbol{v}$, rotation θ
for $t = 0$ to T **do**
    Update mask pose: $\boldsymbol{p}_t \leftarrow \boldsymbol{p}_{t-1} + \boldsymbol{v}, \quad \theta_t \leftarrow \theta_{t-1} + \Delta\theta$
    Handle boundary collisions and bounce $\boldsymbol{v}$ if needed
    Generate binary mask $\boldsymbol{m}_t$ from $S, \boldsymbol{p}_t, \theta_t$
    Extract patch from $\boldsymbol{y}_t$ (using ping-pong loop)
    $\tilde{\boldsymbol{x}}_t \leftarrow \boldsymbol{m}_t \odot \boldsymbol{y}_t + (\mathbf{1} - \boldsymbol{m}_t) \odot \boldsymbol{x}_t$
end for
return $\tilde{\mathcal{X}}$

Algorithm 2 Anchored Control Pipeline (Degraded Reference)

Require: Target video $\mathcal{X}$, Indices set $\{0, \ldots, T\}$
Ensure: Degraded reference video $\hat{\mathcal{X}}$
$\mathcal{K} \leftarrow \{0, T\}$ ▷ Always include start/end
Sample N random indices from $(0, T)$ into $\mathcal{K}$
Sort $\mathcal{K} = \{k_0, k_1, \ldots, k_{|\mathcal{K}|}\}$
for each keyframe index $k \in \mathcal{K}$ **do**
    $\hat{\boldsymbol{x}}_k \leftarrow \boldsymbol{x}_k$
    if random() < 0.5 **then** ▷ Geometric Degradation
        Apply random Zoom-Stretch affine transform to $\hat{\boldsymbol{x}}_k$
    end if
    if random() < 0.5 **then** ▷ Appearance Degradation
        Define local region $\boldsymbol{b}_k$ (random blob)
        $\hat{\boldsymbol{x}}_k \leftarrow (\mathbf{1} - \boldsymbol{b}_k) \odot \hat{\boldsymbol{x}}_k + \boldsymbol{b}_k \odot \text{Blur}(\hat{\boldsymbol{x}}_k)$
    end if
end for
for $t = 0$ to T **do** ▷ Temporal Interpolation
    Find neighbors $k_i, k_{i+1} \in \mathcal{K}$ such that $k_i \le t \le k_{i+1}$
    $\alpha \leftarrow (t - k_i)/(k_{i+1} - k_i)$
    $\hat{\boldsymbol{x}}_t \leftarrow (1 - \alpha)\hat{\boldsymbol{x}}_{k_i} + \alpha\hat{\boldsymbol{x}}_{k_{i+1}}$
end for
return $\hat{\mathcal{X}}$

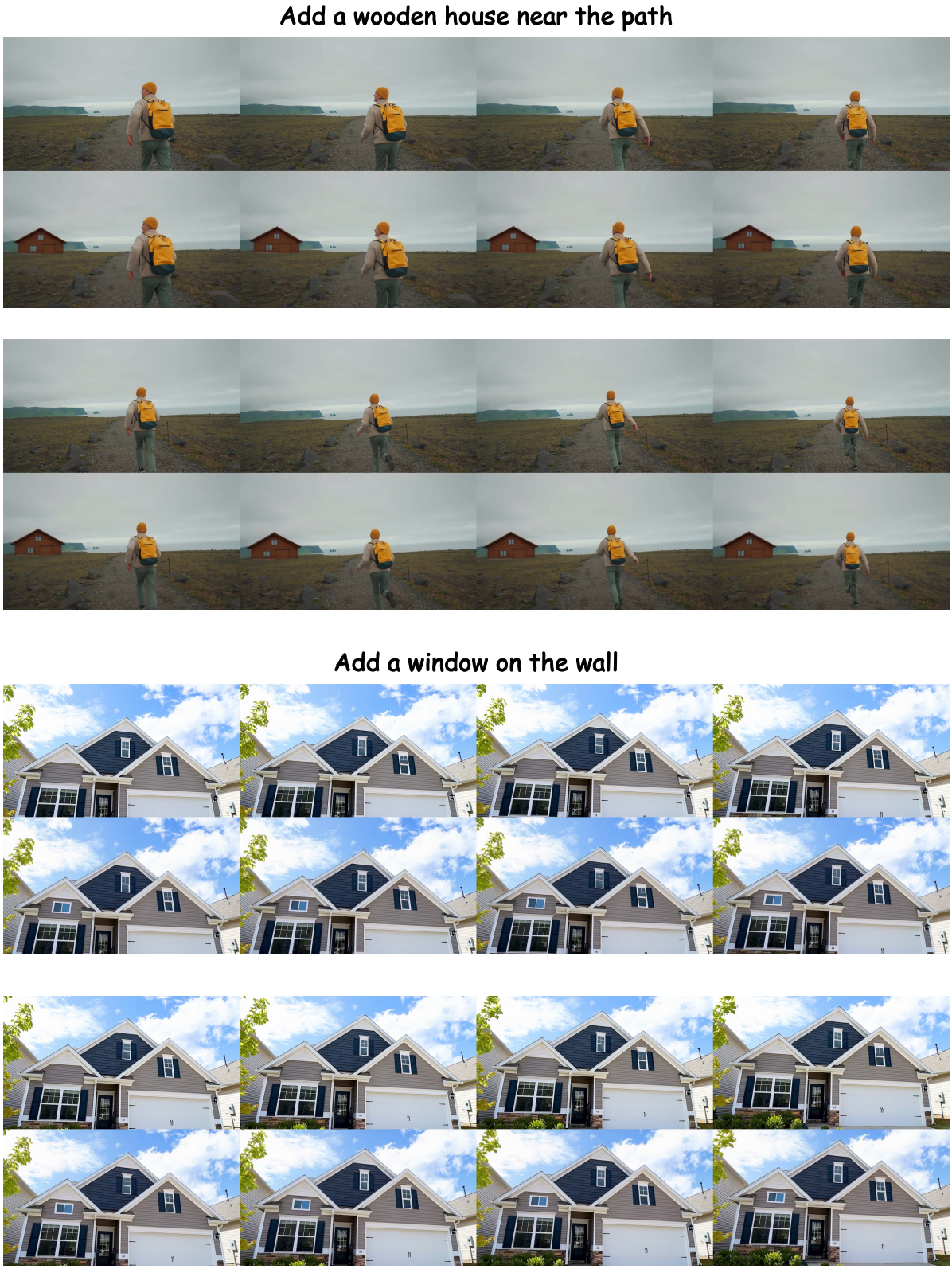

Figure 12. More visual results (1).

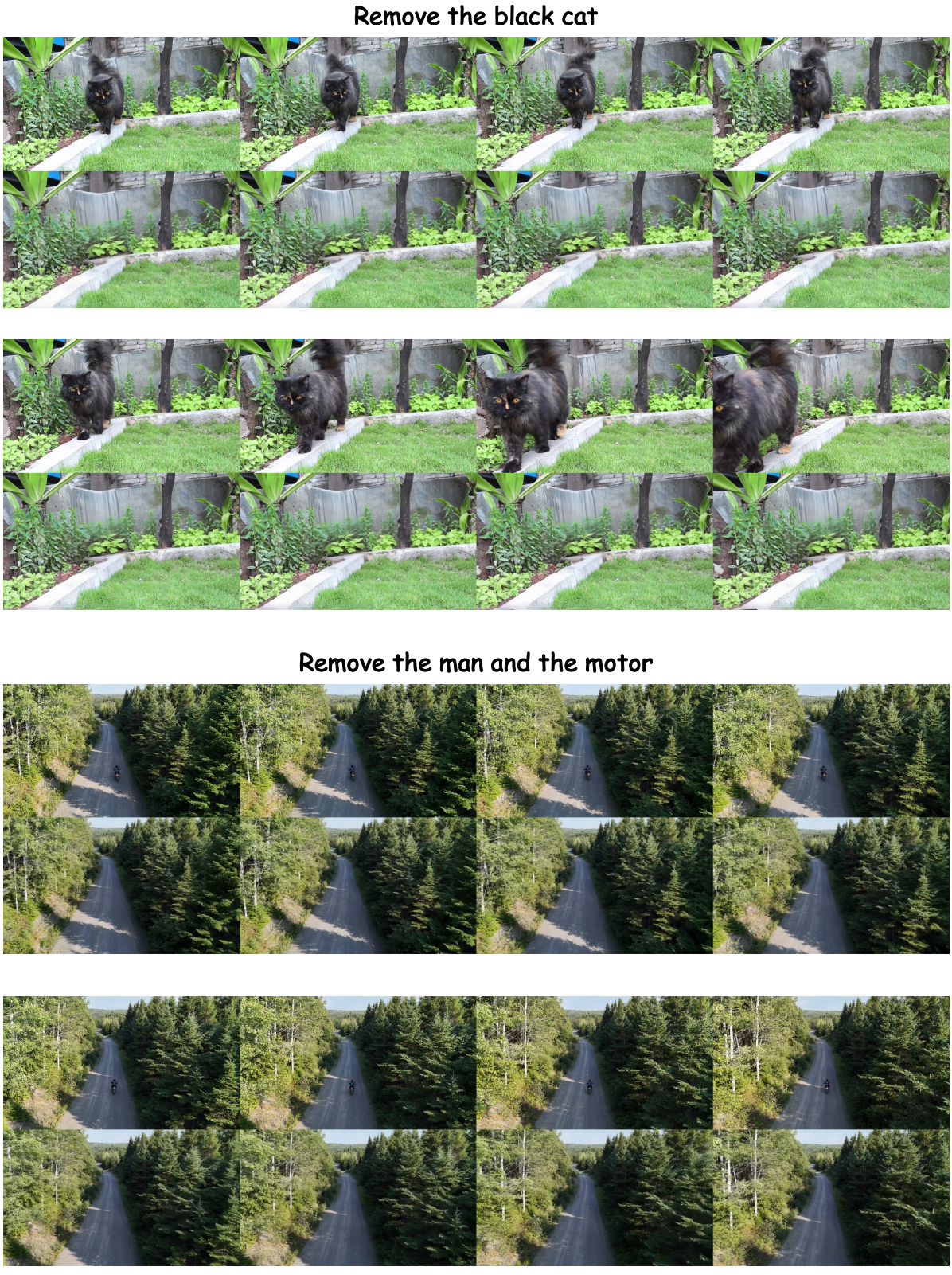

Figure 13. More visual results (2).

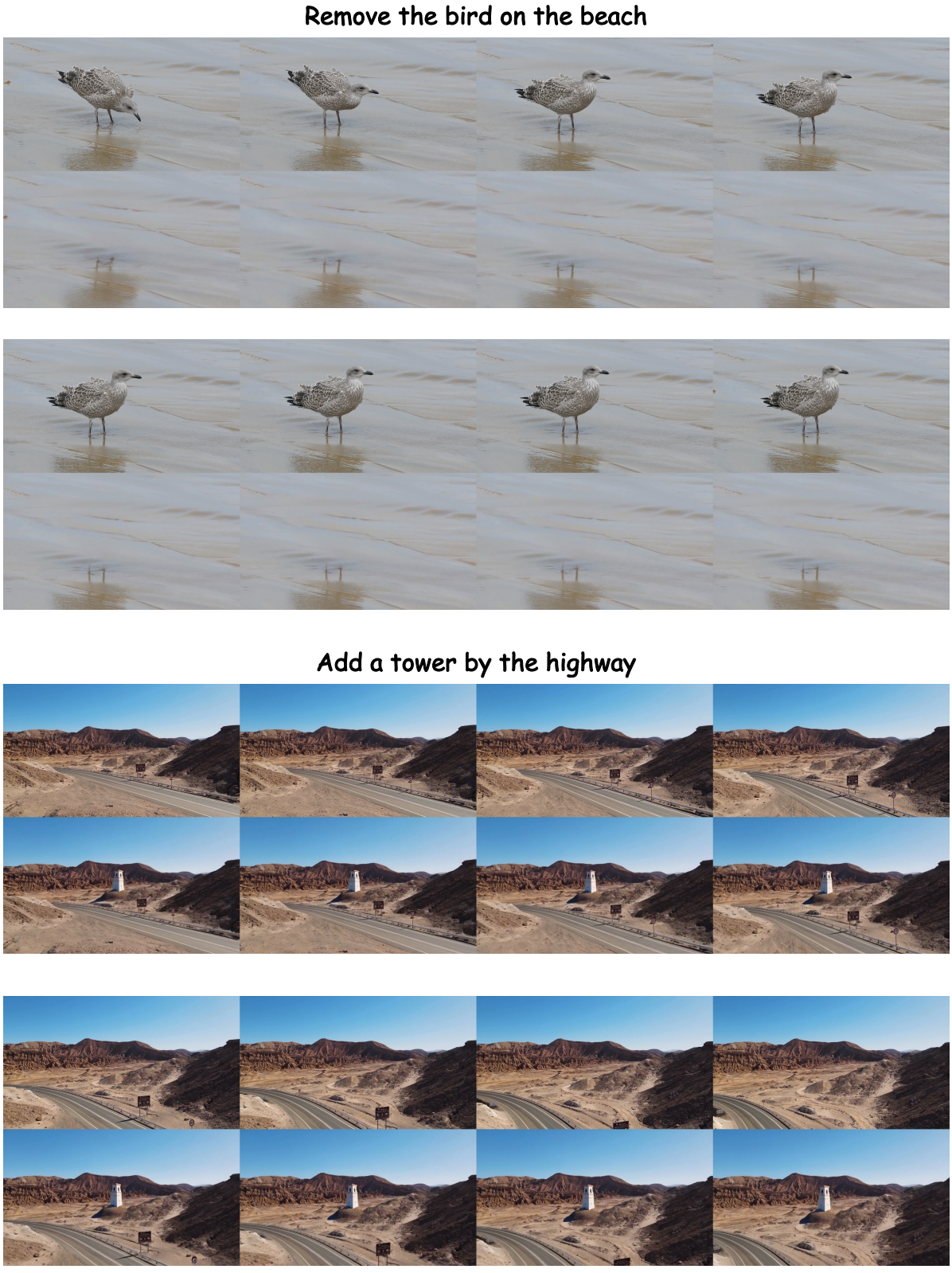

Figure 14. More visual results (3).

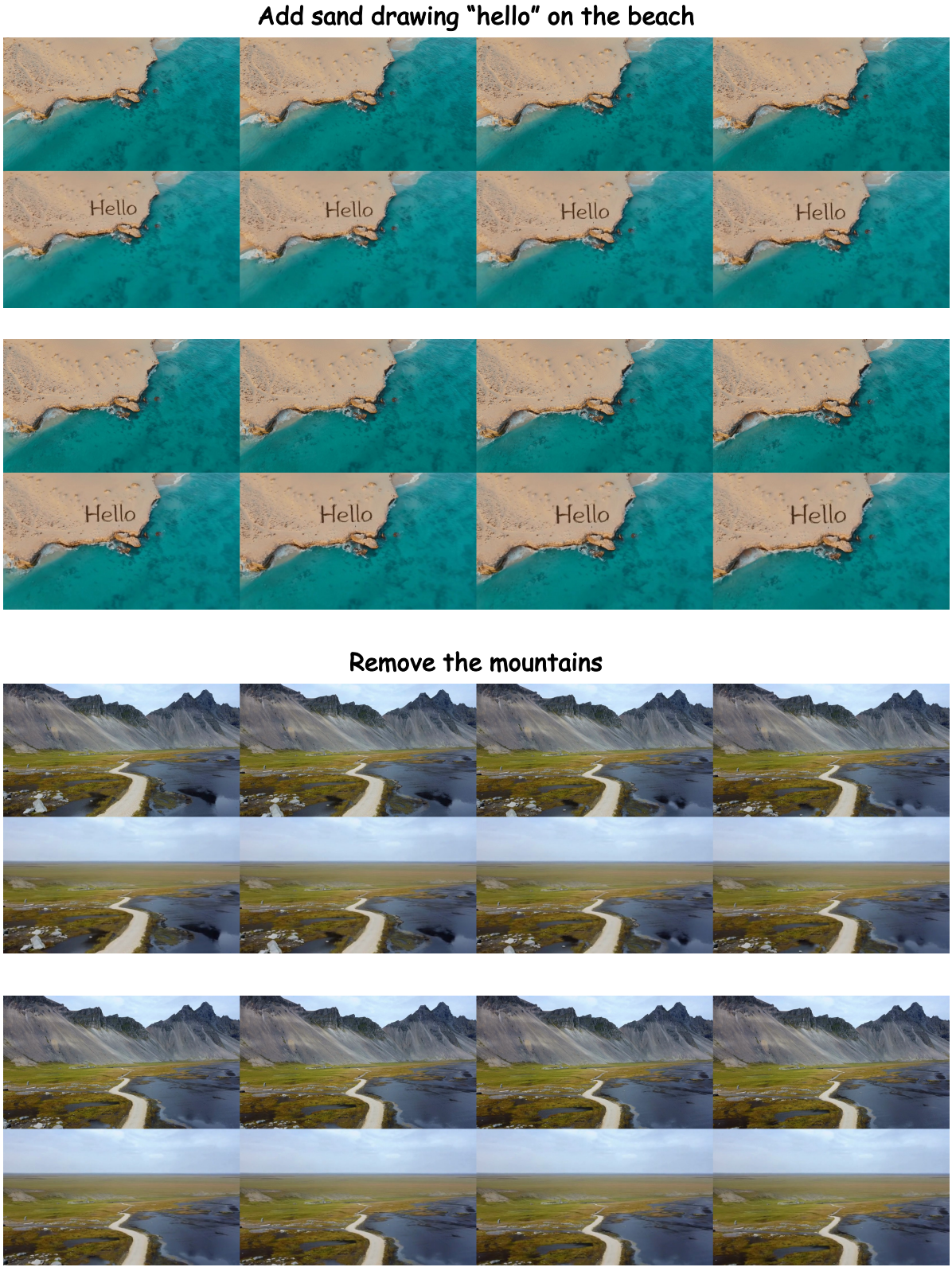

Figure 15. More visual results (4).